社会化问答平台提问回复率的预测研究
——以"百度知道"为例

■ 邓胜利　付少雄　刘瑾
武汉大学信息资源研究中心　武汉 430072

摘要：［目的/意义］基于社会化问答平台提问回复率较低的现状，通过预测提问回复率，能够为社会化问答平台提升用户活跃度与留存率，改善用户体验提供参考。［方法/过程］以"百度知道"为研究平台，抓取平台设置的14个话题下共10 640条提问记录，从提问特征与提问者特征角度，构建提问回复率影响因素的研究框架。采用二元Logistic回归对影响因素进行数据验证，构建提问回复率的预测模型，对模型准确率进行验证。［结果/结论］社会化问答平台提问回复率研究可改善平台信息服务质量与促进用户知识贡献行为，实验结果验证了研究模型在社会化问答平台提问回复率预测中的有效性。

关键词：社会化问答平台　知识贡献行为　回复率　Logistic回归　预测
分类号：G252
DOI：10.13266/j.issn.0252-3116.2019.10.011

1 引言

社会化问答平台（social question & answer community）是以用户提出、回复与讨论问题为主的互联网平台，兼具网络社交网站（social networking services）、社会化媒体（social media）与知识管理系统（knowledge management systems）等社区属性[1-2]。与传统信息搜寻和共享方式相比，虽然社会化问答平台实现了对答案的自我过滤与严格控制，但仍存在回复率较低等问题。如Yahoo! Answers的癌症板块仅有6%的回复率[3]，Google Answers中不到一半提问得到回复[4]。如何有效提升用户回复意愿，从而提高平台的活跃度成为社交问答平台亟需解决的问题。

社会化问答平台要保障用户活跃度与留存率，必须有效提升用户知识贡献意愿，着力于提高平台提问回复率。社会化问答平台回复率相关研究主要集中在两个方面：一是探究平台用户回复的内外在动机[5-10]；二是网络用户信息行为的预测研究，如最佳回答者、转发行为、电影评分行为等[11-17]。经过文献调研，现有研究缺乏对社会化问答平台的回复率预测研究，而对回复率的预测研究有助于改进社会化问答平台。

为解决上述局限性，笔者在对社会化问答平台中提问回复率的影响因素分析的基础上，从提问特征和提问者特征角度，指出提问的载体丰富度、财富值、紧迫性表达与礼貌性表达，及提问者的可靠性、影响力与网络中心度会影响提问回答率。通过python语言爬取"百度知道"下14个话题共10 640条提问记录，采用二元Logistic回归对影响因素进行数据验证，最终构建提问回复率的预测模型，并对模型准确率进行验证。

1 相关研究

1.1 社交问答平台用户回复动机研究

鉴于社会化问答平台中提问的低回复率[3-4]，大量研究对用户贡献知识的内外在动机进行了探究，大都采用问卷或访谈等方法进行实证研究（见表1）。内在动机是指用户完成某些行为所希望获取的积极心理感受，外在动机是指用户渴望在参与某些任务后获得的工具或物质性的奖励[18]。在社会化问答平台中，用户回答问题的内外动机更多是对自身价值的体现。其中，内在动机为用户通过自主选择或自身喜好参与问答，即当用户希望获得采纳或点赞，增加认同感时会参与问答；外在动机主要体现在用户希望获取经验值或财富值等时采取的知识贡献行为[6]。

作者简介：邓胜利（ORCID：0000-0001-7489-4439），教授，博士，博士生导师；付少雄（ORCID：0000-0002-5166-3141），博士研究生，通讯作者，E-mail：fu_shaoxiong@163.com；刘瑾（ORCID：0000-0003-4345-2157），硕士研究生。
收稿日期：2018-11-02　修回日期：2019-01-21　本文起止页码：97-105　本文责任编辑：徐健

表 1　用户回复的内外在动机

动机	平台	结论	文献来源
外在动机	Yahoo！Answers	当提问者对提问给出一定悬赏时，其他用户贡献知识会更加主动且回答质量会更高。	E. Choi, V. Kitzie, C. Shah, 2013[5]
	百度知道	内外在动机共同促进着用户知识贡献行为，其中外在动机影响更大。	徐鹏，张聃，2018[6]
	39 健康问答等	同情、利他主义和互惠对用户知识贡献行为具有显著影响，社会资本中的社会交互和社会信任也会显著提升用户回复意愿。	陈星，张星，曾淑云，2017[7]
	Yahoo！Answers	总结包括享乐、自我效能、学习、个人收益、利他主义、平台兴趣、社交参与、移情、声誉和互惠在内的 10 种动机，同时指出利他主义是影响用户最主要动机，其次是享乐和自我效能。	S. Oh, 2012[8]
内在动机	Yahoo！Answers	提问者观点、问题类型与问题难易程度会影响回复率。	D. Dearman, K. N. Truong, 2010[9]
	Naver	问答平台用户回复的主要驱动力是学习新知识、帮助他人或作为业余爱好参与平台问答活动。	K. K. Nam, M. S. Ackerman, L. A. Adamic, 2009[10]

1.2　网络用户信息行为预测研究

由于当前缺乏对社会化问答平台提问回复率的预测研究，笔者对网络用户信息行为预测研究进行探究，如表 2 所示。用户信息行为预测的研究对象主要集中于社交服务网站和购物网站，根据用户行为的特征数据来构建预测模型。对于预测方法，主要包括 Logistic 回归[19]、随机森林[20]、复杂网络分析[21]、模糊神经网络[22]、决策树[23]等。相关研究预测准确率集中于 50% –94% 之间[11-16]。基于对用户行为预测方法的综合分析，由于提问回复率是典型的二分类问题，Logistic 回归模型是对二分类因变量进行回归分析时应用最普遍的多元量化分析方法，因将目标概率进行 Logit 变换而得以避免线性概率模型的结构缺陷[15]。因此，笔者采用二元 Logistic 回归对社会化问答平台提问回复率进行预测[24]。

表 2　网络用户行为预测研究

平台	预测主题	研究结论	文献来源
百度知道	最佳回答者	基于用户社交网络能够有效预测最佳回答者，并主动寻找最佳回答者。	Q. Du, 2015[11]
Yahoo！Answers	答案质量	基于答案相对位置的顺序，使用排序学习模型框架能够对高质量回答进行预测。	徐安滢，吉宗诚，王斌，2017[12]
Naver	答案质量	基于用户回答采纳率和回答关键词长度等特征能提升回答质量预测准确率。	J. Jeon, W. B. Croft, J. H. Lee, 2006[13]
微博	转发行为	使用马尔可夫随机场框架，基于阅读推广和用户转发行为的影响因素能够有效预测微博用户转发行为。	田磊，任国恒，王伟，2017[14]
某购物网站	购买行为	基于用户购买行为和信息浏览行为的特征，能够有效预测用户商品的购买行为。	张鹏翼等[15]
豆瓣	电影评分	基于用户对电影评价的内容，利用情感分析技术能有效构建电影评分预测模型。	杨红丽等[16]

2　模型构建与假设

社会化问答平台用户知识贡献行为动机可分为外在动机和内在动机[6,18]。其中，内在动机通常出于内心满足或利他主义，更多是心理层面感知。不同于内在动机主要通过问卷与访谈开展研究，笔者采用文本分析方法，从外在动机角度探究提问回复率。以百度知道为研究对象，笔者对社会化问答平台中提问回复率的影响因素指标进行提取。指标提取主要分为提问特征（实体特征、非实体特征）和提问者特征（可靠性、影响力、网络中心度）两个方面，其中实体特征可分为载体丰富度与财富值、非实体特征可分为紧迫性表达与礼貌性表达。社会化问答平台中提问回复率的影响因素研究框架如图 1 所示：

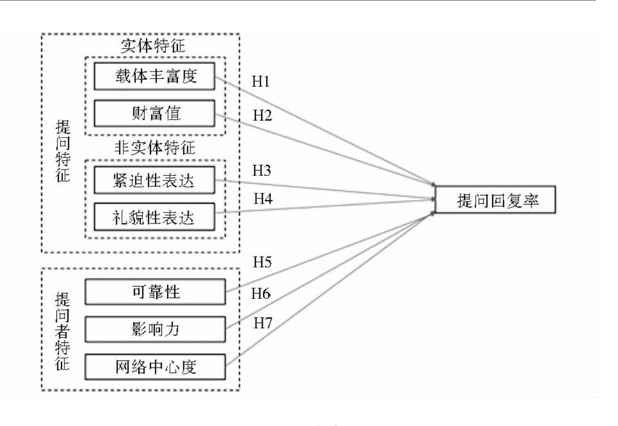


图 1　研究框架

2.1　提问特征

2.1.1　实体特征　实体特征对提问回复率的影响因素包括载体丰富度（career richness）和财富值（wealth value）。

（1）载体丰富度。呈现问题的传播载体（如文本、声音、图形、视频等）会影响用户的视觉和听觉感受[25]。笔者发现在线信息可用纯文本、图片、视频、音频、符号、链接及它们的组合来表达。提问的载体丰富度是指提问内容中使用的表格、图片和外部链接等。包含多种传播载体的内容，用户理解更直观，贡献知识意愿更强。同时，外部链接或图片能够影响答案质量的预测[26]。因此，笔者提出如下假设：

H1：问题内容的载体越丰富，提问回复率越高。

（2）财富值。财富值是社会化问答平台中特有的指标属性，用户可通过签到、回复等行为来获得财富值。同时，提问者可对提问单独设置财富值，被选为最佳答案的回答者可额外获得问题对应的财富值[27]。积累财富值可在平台商城中兑换礼品，同时财富值的增加还可提高平台地位与被认可度。根据社会交换理论，用户倾向于在社会活动中获得更多报酬[6, 28]。付费问题能得到更多回复[4, 29]，在提问中设置财富值会吸引更多用户关注，可能会提升回复率，因此，笔者提出如下假设：

H2：问题设置的财富值越高，提问回复率越高。

2.1.2　非实体特征　非实体特征主要指代情感支持（emotion support），是指提问中悲伤、快乐分享或关心表达的一种形式，可让用户获得温暖，得到关爱与帮助，同时情感支持能积极影响在线社区信息服务的采纳[30]。外部环境的情绪状态能够影响用户决策[31]，而社会问答平台中的情感信息能提升用户的平台认同感[32]，增强信息采纳行为[33]。经过平台调研，提问中非实体特征主要包括紧迫性表达（urgency expression）和礼貌性表达（polite expression）。

（1）紧迫性表达。紧迫性表达是指在提问中表现出立即回复或注意的要求。衡量提问紧迫性包括三个测量指标：紧迫性语句、重复标点、重复感叹词[34]。用户倾向于用紧急词来应对紧急事件，重复性语句来表示传播消息的急切性[34]。在社交问答平台中，用户常用重复标点和感叹词用来强调其焦虑。当提问中有紧迫性表达时，其紧张情绪可能会感染到其他用户，而情绪对于用户行为具有强烈影响[31]，可能会促使用户回复问题。因此，笔者提出如下假设：

H3：提问中的紧迫性表达能积极影响提问回复率。

（2）礼貌性表达。礼貌是用户采取亲近社会行为的主要动机之一[28]。根据社会交换理论，感恩表达能够增强用户自我效能和社会价值，从而鼓励亲近社会行为[28]。礼貌性表达等正向情感信息能增强用户的平台认同感，从而促进用户参与[32]。同时，礼貌性表达会感染回复者情绪[30]，友好态度能提升其成就感，提升其平台交流与回复意愿。基于此，礼貌性表达表达可能提升提问回复率。因此，笔者提出如下假设：

H4：提问中的礼貌性表达能积极影响提问回复率。

2.2　提问者特征

提问者特征主要包括提问者可靠性（questioner reliability）、提问者影响力（questioner impact）与网络中心度（network centrality）。

（1）提问者可靠性。提问者可靠性是其在社会网络中建立信任和增加影响力的重要因素。提问者可靠性一方面可通过用户身份背景确认，另一方面还可通过提问者的专业度来判断，而百度知道中提供的投票机制和声望体系为识别提问者可靠性提供了参考。信息源可靠性可影响该信息被接受的程度，可靠性越低的信息，其被接受程度越低[35]。同时，用户希望能获取可靠知识[36]。因此，提问者可靠性可能会影响提问的被接受程度。因此，笔者提出如下假设：

H5：提问者可靠性越高，问题回复率越高。

（2）提问者影响力。影响力是百度知道中特有的综合评价用户在百度贡献力度、被认可度等方面的指标，每位百度知道用户的个人主页都有其影响力值[37]。根据百度知道官方体系，影响力的影响因素包括用户的回答数、活跃程度、被点赞数、专业认证、作弊情况等。用户影响力越高，其活跃度越高，与其他用户交互更多，在平台中的地位和影响也会更高[38]。具有较大影响力的用户，其提问更有可能被回复。因此，笔者提出如下假设：

H6：提问者影响力越高，问题回复率越高。

（3）网络中心度。提问者的网络中心度包括外向中心度和内向中心度。外向中心度指提问者关注其他用户的数量，内向中心度指提问者的被关注数量。提问者外向中心度越高，其回复行为可能越多；内向中心度越高，其提问可见度越高[39]。同时，平台中好友间的交互性要显著高于陌生人间的交互性，提问者的好友选择回答其提问的可能性也越大[16]。提问者的网络中心度越高，其提问获得回复的概率可能越高。因此，笔者提出如下假设：

H7：提问者网络中心度越高，问题回复率越高。

3　研究方法

3.1　数据来源

百度知道的话题包括“经济金融”“健康生活”“娱

乐休闲”“体育运动”等 14 个话题。利用 Python 语言抓取每个话题下的提问,同时保证每条抓取数据的提问内容是非空的,剔除重复的提问数据并补充非重复的新提问数据,确保每个话题下有效提问数据量的一致性。截止到2017 年 10 月 20 日,共抓取 14 个话题下共 10 640 个提问内容,每个话题 760 个提问内容。然后将原始数据集存储在数据库 MySQL 中,进行数据预处理。针对提问数据,主要记录了提问时间、提问标题、提问内容、提问浏览数、提问财富值、提问者昵称、提问者影响力、提问者历史回答数、提问媒介;针对答案数据,主要记录了答案数量、最佳答案内容、最佳答案时间、其他答案内容、其他答案时间、回答媒介。

3.2 预测方法

笔者首先使用 SPSS 22.0 对数据进行描述性和相关性分析, 从总体资源分布、用户提问和回答行为、高质量回答的角度来分析社会化问答社区中知识贡献行为的特点,以及提问回复率的影响因素。然后,由于在社会化问答平台中,用户面对提问可选择回复或不回复,提问是否得到回复可视为典型的二分类问题。而 Logistic 回归模型是典型非线性的回归分析方法,可对统计分析结果进行分类,得到概率性的预测结果[23]。其中,二元 Logistic 回归是因变量值只为 0 和 1 的 Logistic 回归,是处理二分类问题的常用方法。因此,利用二元 Logistic 回归模型对影响因素进行数据验证,构建提问回复率的预测模型。具体而言,将数据集分为测试集与训练集,随机抽取部分数据作为测试集,用于验证预测模型的准确性,剩余数据作为训练集,对影响因素进行验证并基于数据分析结果构建预测模型,及对模型的正确率进行验证。

3.3 变量测量

基于研究模型与假设,研究包含因变量(回复数)与自变量(载体丰富度、紧迫性表达等)。对于因变量,得到回复标记为 1,未得到回复标记为 0。结合百度知道平台的数据特征,自变量的测量方式如下:

(1)实体特征。对于载体丰富度,提问中包含表格、图片或链接标记为 1,未包含标记为 0;对于财富值,可从提问界面直接获取,提问界面未设置财富值标记为 0,否则值为具体财富值。

(2)非实体特征。对于紧迫性表达,首先编写 Python 程序对提问文本进行句子切割。其次使用分词软件 ICTCLAS 对句子进行分词。然后根据知网的情感词表中体现紧迫性的词语、事先定义的重复性标点符号和没有实际意义的情感词进行自动化查询匹配[40]。句子中若包含上述紧迫性表达,则标注为 1;若不包含,则标注为 0。

对于礼貌性表达,笔者首先预定义包含 14 个相关词的列表,如“谢谢”“感激”“回报”等,然后利用上述列表与分词结果进行查询匹配。若包含礼貌性表达,则标注为 1;若不包含,则标注为 0。

(3)提问者特征。首先通过回答者个人主页获取提问者的影响力值、回答数、帮助人数,以及提问者的关注数和被关注数,主要测量如下指标:①对于提问者可靠性,定义为提问者的回答数除以提问者帮助人数;②对提问者的影响力按照数值进行划分,影响力为 0 时值为 0;影响力为 0 –50 时值为 1;影响力值为 50 –100 时值为 2;影响力为 100 –150 时值为 3;影响力为 150 –200 时值为 4;影响力大于 200 时值为 5;③提问者的网络中心度包含外向中心度和内向中心度,外向中心度指提问者的关注数,内向中心度指提问者的被关注数。

4 研究结果

4.1 描述性统计

(1)总体资源分布。在所有的样本中,抽样问题共 10 640 个,回答 34 048 个,平均一个问题对应 3.2 个回答,不同回答数量下的提问分布如图 2 所示。其中,回答数为 0 的提问占 20.3%,回答数量为 1 个占 25.9%。同时,回答数大于等于 8 个的提问仅占 1.4%,百度知道中获取大量回答的提问占比较低。

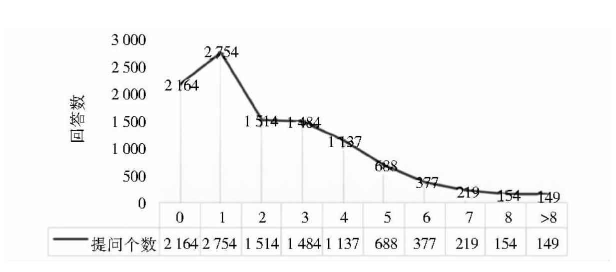


图 2 不同回答数量下的提问分布

(2)回复时间统计。回答时间总体分布如图 3 所示,其中每天 11 点 –14 点和 21 点 –23 点时,用户回复量最高,晚上 21 点达到回答量的最高峰。可能的原因是上述时间段为非工作时间段,用户有时间贡献知识。同时,14 点 –19 点及凌晨以后,回答数量呈明显下降趋势,可能上述时间段为工作与休息时间段。回答时间总体分布见图 3。

(3)提问主题。不同主题类型答案数量如表 3 所示,其中每个主题的问题数皆为 760。医疗卫生类、社会民生类、心理分析类回复数占比最高,分别为

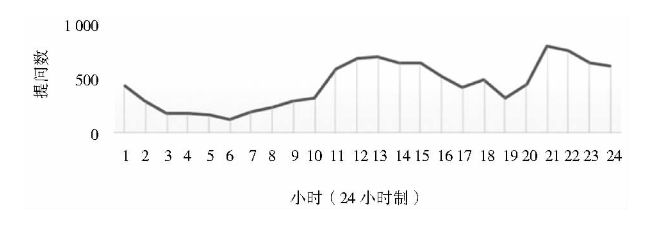


图 3　回答时间总体分布

10.37%、12.74%与 11.88%，用户对上述主题问题关注度较高。相反，企业管理类、娱乐休闲类和行政地区类回复数占比较低，分别为 1.73%、4.32% 和 4.32%。总体上，不同主题问题的回复数存在差异性，主题类型会影响问题回复率。

表 3　不同主题类型答案数量

主题类型	回复数(个)	百分比(%)
电脑网络	3 309	9.72
健康生活	2 720	7.99
医疗卫生	3 531	10.37
体育运动	2 061	6.05
电子数码	2 499	7.34
经济金融	1 546	4.54
科学教育	2 646	7.78
社会民生	4 338	12.74
文化艺术	1 617	4.75
法律法规	2 206	6.48
娱乐休闲	1 471	4.32
心理分析	4 045	11.88
企业管理	5 890	1.73
行政地区	1 471	4.32

(4)客户端和匿名行为。对于客户端，从提问者角度，64.8%的提问来自手机客户端，35.2%的提问来自电脑客户端。相反，只有 38.6%的回复来自手机端，60.4%来自电脑端。可能的原因是，手机客户端更为便捷，提问者倾向于使用手机提问。而回答者需要查询资料进行回复，更习惯使用电脑作答。

对于匿名行为，提问中 29.1%用户是匿名的，70.9%是非匿名的；而回复中 20.8%的回答是匿名的，79.2%是非匿名的。选择匿名提问的用户要稍微多于匿名回答的用户，可能的原因是提问会涉及用户私人、身体状况等敏感信息，提问者为保护隐私选择匿名提问，而回答者为获得其他用户认可，而选择非匿名方式。客户端与匿名统计具体见表 4。

4.2　预测模型构建及假设检验

4.2.1　预测模型的构建　根据前文分析，29.1%用户选择匿名发布提问，此类用户无法获取到点赞数、影响值等信息，也无法获取其提问是否得到回答，因此匿名

表 4　客户端与匿名统计

属性	客户端		匿名		
		手机	电脑	匿名	不匿名
提问	频数	6 895	3 745	3 096	7 544
	百分比(%)	64.8	35.2	29.1	70.9
回复	频数	4 107	6 533	2 213	8 427
	百分比(%)	38.6	61.4	20.8	79.2

数据不作为构建预测模型的样本数据。剔除上述匿名数据，共有 7 544 条有效提问数据。随机抽取 200 条数据作为测试集，用来检验预测模型的准确性，其余 7 344 条数据作为训练集，用以构建提问回复率的预测模型。

笔者使用 SPSS statistics 22.0 中的二元 Logistics 回归分析模块构建预测模型，并对模型参数进行评估及显著性检验。将变量值按照测量指标的计算方式，进行整理并计算出模型中所涉及的各变量值，选择二元 Logistic 回归完成模型训练，模型变量参数及各项统计指标如表 5 所示：

表 5　预测模型中方程式的变量

指标	B	S. E	Wald	Df	Sig.	Exp(B)
载体丰富度	46.593	17.1191	4.048	1	.011	1.051E+15
财富值	14.589	18.949	6.046	1	.014	1.718E+20
紧迫性表达	16.057	15.081	5.716	1	.017	4.562
礼貌性表达	12.230	1.397	2.541	1	.003	9.296
提问者可靠性	.285	.059	22.933	1	.044	1.329
提问者影响力	.399	1.377	0.061	1	.806	1.403
提问者网络中心度	1.021	.277	13.559	1	.000	2,775
常量	－22.797	10.379	4.824	1	.028	.000

4.2.2　模型参数及假设检验　在 Logistic 回归模型建立后，首先对模型拟合度效果进行检验，其中－2 Log likelihood 值用于检验模型的整体拟合效果，该值为 20.832，大于卡方临界值 5.991，该模型拟合效果较好。此外，Cox-Snell 拟合优度和 Nagelkerke 拟合优度值越接 1，拟合度越好，结果也表明该模型拟合度较好；然后开展模型系数的综合检验，由模型系数的 Omnibus 检验发现，步骤、模块和模型的卡方值都大于临界值 5.991，显著性要远小于临界值 0.05，模型系数检验通过；最后对 Hosmer-Lemeshow 拟合优度进行检验，当卡方统计值＜卡方临界值，sig. ＞0.05 时，接受假设。结果显示模型能很好拟合整体，模型预测值与观测值不存在显著差异，具体见表 6。

在此基础上，笔者对假设进行检验，Logistic 回归检验结果显示：①H1 成立，当提问中使用外部链接、图片、表格时，回复率更高。提问中的外部链接、图片、表

表 6　模型参数

模型摘要			
Step	-2 Log likelihood（-2 对数似然）	Cox-Snell R square（考克斯-斯奈尔 R 方）	Nagelkerke R square（内戈尔科 R 方）
	20.832	.652	.809
模型系数的 omnibus 检验			
	Chi-square（卡方）	df	.sig
Step	29.187	2	.000
Block	29.187	2	.000
Model	29.187	2	.000
Hosmer-lemeshow 拟合优度检验			
Step	Chi-square（卡方）	df	.sig
	4.739	9	.583

格能提升用户的视觉感官，从而提升用户理解提问内容的效率，以及提升用户回复率[25]；②H2 成立，对提问设置一定财富值，对回答者吸引力更大，回复率更高。用户在包括社会化问答在内的社会活动中倾向于获得更多的报酬[6,28]，而财富值是社会化问答平台中报酬的体现，能够提升用户平台中的被认可度，所以财富值的提供可以提升提问回复率[4,27,29]；③H3 成立，当提问描述中包含紧迫性的词语或符号时，回复率更高。紧迫性的词语或符号能够表达提问内容的紧迫性，进而影响其他用户的回复行为[31,34]；④H4 成立，当提问描述中包含礼貌性的词语时，回复率更高。礼貌性的表达能增强用户间的亲近感[28]，而且礼貌性等正向情感表达可积极影响回答者情绪，提升用户的回复参与度[30, 32]；⑤H5 与 H6 不成立，提问者可靠性与影响力并未对回复率产生显著影响。在社会化问答平台中，提问内容首页仅显示用户头像与昵称等信息，用户需进入提问者主页才能分析其影响力值，所以用户在回复提问中较多关注提问内容本身，较少关注提问者的可靠性与影响力，从而对回复率的影响较小；⑥H7 成立，提问者的用户关注数和被关注量较高时，提问回复率较高。提问者的用户关注数和被关注量越高，提问者的网络中心度也越强，与其他用户的交互度则越强[39-40]。所以用户提问内容的曝光度越高，对应的提问回复率也能得到提升。

4.3　预测模型验证

4.3.1　实验过程　根据模型参数及假设检验结果，可得如下 Logistic 回归方程，其中 X1 为载体丰富度；X2 为财富值；X3 为紧迫性表达；X4 为礼貌性表达；X5 为网络中心度。

$$P(Y=1 \mid X)=f(X)=\frac{1}{1+e^{-g(X)}}=\frac{1}{1+e^{-(-22.797+49.593X1+14.589X2+16.057X3+12.230X4+1.021X5)}}$$

笔者根据回归方程计算观测值概率并进行预测，当概率小于 0.5 时，预测结果为提问未被回答；概率值大于 0.5 时，预测结果为提问至少得到 1 个回答。利用 200 条测试集数据计算模型中各变量值，其中“观察是否得到回答”表示实际是否得到回答，“预测是否得到回答”表示根据各观测组的自变量值，预测是否得到回答。“预测结果”通过“观测是否得到回答”和“预测是否回答”对比所得结果。

4.3.2　预测模型评价　经过多次迭代运算，回归模型的各参数逐渐收敛并稳定，从而得到最终模型参数和指标。基于最终 Logistic 回归模型，对提问回复率进行预测。预测结果分类表如表 7 所示，根据分类表构建预测准确率、灵敏度、特异度、漏诊率与误诊率的相关公式，见表 8。

表 7　预测结果分类

属性	预测得到回答	预测未得到回答
实际得到回答	a	b
实际未得到回答	c	d

表 8　评价指标公式

评价指标	指标描述	公式
预测准确率	表示实际分类 y = 1（得到回答）的个体中预测结果也为 1（得到回答）和实际分类 y = 0（没有得到回答）的研究样本中预测结果也为 0（没有得到回答）的概率	$P=\frac{a+b}{a+b+c+d}=100\%$
灵敏度	指实际分类 y = 1（得到回答）的研究样本中预测结果也为 1（得到回答）的概率	$TPR=\frac{a}{a+b}\times 100\%$
特异度	指实际分类 y = 1（得到回答）的研究样本中预测结果却为 0（没有得到回答）的概率	$TNP=\frac{d}{c+d}\times 100\%$
漏诊率	指实际分类 y = 1（得到回答）的研究样本中预测结果却为 0（没有得到回答）的概率	$FNR=1-TPR=\frac{b}{a+b}\times 100\%$
误诊率	指实际分类中 y = 0（没有得到回答）的研究样本中预测结果却为 1（得到回答）的概率	$FPR=1-TNR=\frac{c}{c+d}\times 100\%$

通过测试集计算，测试集预测准确率为 91.1%，社会化问答平台提问回复率的预测准确率达较高水平[11-16]。同时，实际得到回答而预测未得到回答的占比 9.5%，漏诊率为 9.5%。实际未得到回答而预测得

到回答的占比8.3%,误诊率为8.3%。但总体上,笔者提出的预测模型准确率能较好预测提问回复率。同时,ROC曲线能较好评价二分类问题的分类效果。预测模型的ROC曲线如图4所示,其中ROC图形中曲线下方面积越大,则模型预测效果越好。曲线下方面积(AUC)的各项检验值曲线下方面积占比0.693,显著性在0.05水平下显著,预测模型可较好预测提问回复率。见表9。

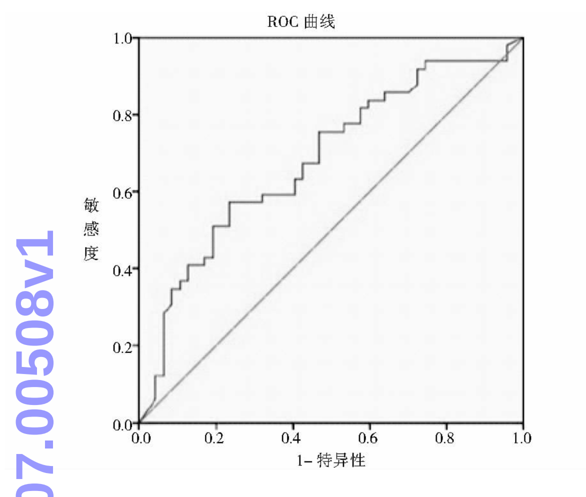


图4 预测模型ROC示意

表9 ROC图曲线下面积

面积图	标准错误	渐进显著性水平	渐进95%置信区间	
			下限值	上限值
0.693	0.55	0.002	0.556	0.793

5 结语

社会化问答平台提问的低回复率既无法有效满足用户信息需求,也不能促进平台的可持续性发展。笔者通过Python编写爬虫实现百度知道数据的采集,利用二元logistic回归构建提问回复率预测模型,并对模型准确率进行验证,研究结果显示笔者提出的模型能较好预测社会化问答平台提问回复率。

笔者探究了社会化问答平台中低回答率的问题,可为平台信息服务改善与知识贡献行为改善提供参考:对于社会化问答平台,对提问回复率进行预测,可为“百度知道”“知乎”“Yahoo! answers”等问答平台的策略制定,提问率提升给予借鉴。如参考高回复率问题的提问特征,平台可为用户提升回复率提供有针对性的建议;通过预测回答率较低的问题,重点加强此类问题的个性化推送,将问题推荐给相关领域专家,以提升回复率;分析与识别可能影响社会化问答平台提问回答率的外部因素,可将其纳入社会化问答工具或服务设计中。对于社会化问答平台用户,可着力于自身可靠性、影响力与网络中心度的提升。同时,依据回答率较高的提问方式,完善提问技巧,如增强提问的载体丰富度等。

同时,本文也存在研究局限性:首先是研究变量,笔者仅从外在动机角度进行探究,未探究内在动机,未来可结合问卷和访谈等方法对内在动机的测量[41-42],分析内外在动机的交互作用对回复率的影响;其次是研究对象,本文的研究对象“百度知道”,属于第1代社会化问答平台,用户间社交联系较弱。未来研究可拓展研究对象,探究预测模型在知乎等第2代社会化问答平台上的适用性;再者,后续可结合问答平台中不同板块进行实证研究,分析不同主题问答的预测机制;此外,本文提问的回答数量是以回答数和未回答数来作为因变量,未来回复率研究可考虑1条或多条回答的差异。

参考文献:

[1] 付少雄,陈晓宇,邓胜利.社会化问答社区用户信息行为的转化研究——从信息采纳到持续性信息搜寻的理论模型构建[J].图书情报知识,2017(4):80-88.

[2] 陈晓宇,付少雄,邓胜利.社会化问答用户信息搜寻的影响因素研究——一种混合方法的视角[J].图书情报工作,2018,62(20):102-111.

[3] ADAMIC L A, ZHANG J, BAKSHY E, et al. Knowledge sharing and yahoo answers: everyone knows something[C]// Proceedings of the 17th international conference on world wide web. New York: ACM, 2008: 665-674.

[4] RAFAELI S, RABAN D, RAVID G. How social motivation enhances economic activity and incentives in the google answers knowledge sharing market[J]. Social science electronic publishing, 2007, 3(1):1-11.

[5] CHOI E, KITZIE V, SHAH C. "10 Points for the best answer!" - baiting for explicating knowledge contributions within online Q&A[C]// Proceedings of the 76th ASIS&T annual meeting: beyond the cloud: rethinking information boundaries. Montreal: Wiley, 2013: 103-106.

[6] 徐鹏,张聃.网络问答社区知识分享动机探究——社会交换论的视角[J].图书情报知识,2018(2):105-112.

[7] 陈星,张星,曾淑云.健康问答平台中知识分享意愿的影响因素研究[J].现代情报,2017,37(4):62-71.

[8] OH Sanghee. The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments[J]. Journal of the Association for Information Science & Technology, 2012, 63(3):543-557.

[9] DEARMAN D, TRUONG K N. Why users of Yahoo! answers do not answer questions[C]// Proceedings of the 2010 SIGCHI con-

ference on human factors in computing systems. New York: ACM, 2010: 329 – 332.
[10] NAM K K, ACKERMAN M S, ADAMIC L A. Questions in, knowledge in: a study of naver's question answering community [C]// Proceedings of the 2009 international conference on human factors in computing systems. New York: ACM, 2009: 779 – 788.
[11] DU Qing. A relationship-based social question and answer system in social network[J]. Journal of information & computational science, 2015, 12(10):3783 – 3798.
[12] 徐安滢, 吉宗诚, 王斌. 基于用户回答顺序的平台问答答案质量预测研究[J]. 中文信息学报, 2017, 31(2):132 – 138.
[13] JEON J, CROFT W B, LEE J H, et al. A framework to predict the quality of answers with non-textual features[C]//Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2006: 228 – 235.
[14] 田磊, 任国恒, 王伟. 面向阅读推广的微博用户转发行为预测[J]. 情报学报, 2017(11):1175 – 1182.
[15] 张鹏翼, 王丹雪, 焦祎凡,等. 基于用户浏览日志的移动购买预测研究[J]. 数据分析与知识发现, 2018, 2(1):51 – 63.
[16] 张红丽, 刘济郢, 杨斯楠,等. 基于网络用户评论的评分预测模型研究[J]. 现代图书情报技术, 2017, 1(8):48 – 58.
[17] 邓胜利, 付少雄. 网络谣言特征分析与预测模型设计:基于用户信任视角[J]. 情报科学, 2017(11):8 – 12.
[18] 陈君, 何梦婷. 基于动机视角的虚拟社区即时/持续网络口碑传播研究[J]. 情报科学, 2017(11):126 – 131.
[19] 陈姝, 窦永香, 张青杰. 基于理性行为理论的微博用户转发行为影响因素研究[J]. 情报杂志, 2017, 36(11):147 – 152.
[20] 安璐, 易兴悦, 余传明,等. 突发公共卫生事件微博影响力的预测研究[J]. 情报理论与实践, 2017, 40(8):76 – 81.
[21] 王林森, 王学义. 微博内向型传导热点发现与预测算法研究[J]. 图书情报工作, 2018,62(3):71 – 77.
[22] 胡悦, 王亚民. 基于模糊神经网络的微博舆情趋势预测方法[J]. 情报科学, 2017, 35(12):28 – 33.
[23] AGICHTEIN E, CASTILLO C, DONATO D. Finding high-quality content in social media[C]// Proceedings of the 2008 international conference on web search and data mining. New York: ACM, 2008:183 – 194.
[24] FLEISCHMANN M, BLOEMHOF-RUWAARD J M, DEKKERB R, et al. Quantitative models for reverse logistics: a review[J]. European journal of operational research, 1997, 103(1):1 – 17.
[25] MCMILLAN S J. Effects of Structural and perceptual factors on attitude toward the website[J]. Journal of advertising research, 2004, 43(4): 400 – 421.
[26] TIAN Q, ZHANG P, LI B. Towards predicting the best answers in community-based question-answering services[C]// Proceedings of the seventh international AAAI conference on weblogs and social media. Palo Alto: AAAI Press, 2013: 725 – 728.
[27] 百度知道. 财富值[EB/OL]. [2018 – 07 – 30]. http://help.baidu.com/question? prod_id = 9&class = 338&id = 1506.
[28] HOMANS G C. Social behavior as exchange[J]. American journal of sociology, 1958, 63(6):597 – 606.
[29] HSIEH G, KRAUT R E, HUDSON S E. Why pay?: exploring how financial incentives are used for question & answer[C]// Proceedings of the 2010 SIGCHI conference on human factors in computing systems. Atlanta: ACM, 2010: 305 – 314.
[30] 吴江, 李姗姗. 在线健康社区用户信息服务使用意愿研究[J]. 情报科学, 2017(4):119 – 125.
[31] 徐健. 基于网络用户情感分析的预测方法研究[J]. 中国图书馆学报, 2013, 39(3):96 – 107.
[32] JOYCE E, KRAUT R E. Predicting continued participation in newsgroups[J]. Journal of computer-mediated communication, 2006, 11(3): 723 – 747.
[33] 金家华. 社会化问答平台中用户知识行为的影响因素研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.
[34] BALDWIN C L, MAY F. Verbal collision avoidance messages of varying perceived urgency reduce crashes in high risk scenarios [C]// Proceedings of the third international driving symposium on human factors in driver assessment, training and vehicle design. Rockport: Public Policy Center of University of Iowa, 2017: 128 – 133.
[35] 沈旺, 国佳, 李贺. 网络社区信息质量及可靠性评价研究——基于用户视角[J]. 数据分析与知识发现, 2013, 29(1):69 – 74.
[36] HEIM J. Why people use social networking sites[C]// Proceedings of the 3rd international conference on online communities and social computing. Berlin: Springer, 2009:143 – 152.
[37] 百度知道. 用户影响力[EB/OL]. [2018 – 07 – 30]. https://jingyan.baidu.com/article/84b4f5659102b460f6da322c.html.
[38] 彭丽徽, 李贺, 张艳丰. 基于灰色关联分析的网络舆情意见领袖识别及影响力排序研究——以新浪微博"8·12 滨海爆炸事件"为例[J]. 情报理论与实践, 2017, 40(9): 90 – 94.
[39] WENG J, LIM E P, JIANG J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]// Proceedings of the third ACM international conference on web search and data mining. New York: ACM, 2010: 261 – 270.
[40] 知网. 情感分析用词语集(beta 版)[EB/OL]. [2018 – 07 – 30]. http://www.keenage.com/html/c_index.html.
[41] 邓胜利, 付少雄. 社交媒体附加信息对用户信任与分享健康类谣言的影响分析[J]. 情报科学, 2018, 36(3): 51 – 57.
[42] 付少雄, 胡媛. 大学生健康信息行为对实际健康水平的影响研究——基于健康素养与健康信息搜寻视角[J]. 现代情报, 2018, 38(2): 84 – 90, 105.

作者贡献说明:
邓胜利:提出论文思路,修订论文最终版本;
付少雄:设计研究框架,起草和修改论文;
刘瑾:数据收集与分析。

The Prediction Research of Response Rate in Social Q&A Communities: A Case Study of Baidu Knows

Deng Shengli　Fu Shaoxiong　Liu Jin
Center for Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] Based on the current situation of low response rate of social Q&A communities, the research can provide references for social Q&A communities to improve user activation, retention rate and user experience by predicting the response rate of questions. [**Method/process**] The paper took "Baidu Know" as the research platform, and grabbed 10 640 question records under 14 topics set by the platform. From the perspective of question and questioner characteristics, the paper constructed the research framework of the factors affecting the question response rate. The binary logistic regression was used to verify the influencing factors, and then the prediction model of the question response rate was constructed. [**Result/conclusion**] The prediction research of response rate in social Q&A communities can improve the quality of platform information services and promote user knowledge contribution behavior. The experimental results have verified the validity of the model in the prediction of question response rate of the social Q&A communities.

Keywords: social Q&A community　knowledge contribution behavior　response rate　logistic regression　prediction

